Emergent and Predictable Memorization in Large Language Models

Year: 2023 | Citations: 159 | Authors: Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika

Abstract

Memorization, or the tendency of large language models (LLMs) to output entire sequences from their training data verbatim, is a key concern for safely deploying language models. In particular, it is vital to minimize a model's memorization of sensitive datapoints such as those containing personal identifiable information (PII). The prevalence of such undesirable memorization can pose issues for model trainers, and may even require discarding an otherwise functional model. We therefore seek to predict which sequences will be memorized before a large model's full train-time by extrapolating the memorization behavior of lower-compute trial runs. We measure memorization of the Pythia model suite and plot scaling laws for forecasting memorization, allowing us to provide equi-compute recommendations to maximize the reliability (recall) of such predictions. We additionally provide further novel discoveries on the distribution of memorization scores across models and data. We release all code and data necessary to reproduce the results in this paper at https://github.com/EleutherAI/pythia